Estimating Common Principal Components in High Dimensions

Ryan P. Browne* and Paul D. McNicholas

Department of Mathematics & Statistics, University of Guelph.

Abstract

We consider the problem of minimizing an objective function that depends on an orthonormal matrix. This situation is encountered when looking for common principal components, for example, and the Flury method is a popular approach. However, the Flury method is not effective for higher dimensional problems. We obtain several simple majorization-minizmation (MM) algorithms that provide solutions to this problem and are effective in higher dimensions. We then use simulated data to compare them with other approaches in terms of convergence and computational time.

1 Introduction

The minimization of the objective function

$$f(\mathbf{D}) = \sum_{g=1}^{G} \text{tr}\{\mathbf{W}_g \mathbf{D} \mathbf{A}_g^{-1} \mathbf{D}'\} \tag{1}$$

is required for a potpourri of statistical problems. To minimize an objective function $f(\mathbf{D})$ that depends on a orthonormal matrix $\mathbf{D}$ (such as an eigenvector matrix), the search space is the orthogonal Stiefel manifold. This manifold is an embedded sub-manifold of $\mathbb{R}^{p\times p}$ equal to the set of all orthonormal matrices $\mathcal{M} = \{\mathbf{D} \in \mathbb{R}^{p\times p} : \mathbf{D}'\mathbf{D} = \mathbf{I}_p\}$, where $\mathbf{I}_p$ denotes the p-dimensional identity matrix. The matrices $\mathbf{W}_1, \ldots, \mathbf{W}_G$ are positive-definite and are usually sample covariance matrices. The matrices $\mathbf{A}_1, \ldots, \mathbf{A}_G$ are diagonal matrices with positive elements.

In Flury and Gautschi (1984) a common principal components model for G groups is found by minimizing (1). Schott (1998) and Boik (2003) use this objective function to find common principal components on correlation matrices. Merbouha and Mkhadri (2004) use this objective function as a regularized technique in discriminant analysis with mixed discrete and continuous variables for generalized location models. Yang and Shahabi (2006) use this decomposition for multivariate time series data sets and use it in various multimedia, medical and nancial applications. Celeux and Govaert (1995) give an expectation-maximization (EM) algorithm (Dempster et al., 1977) wherein each M-step the minimization (1) is preformed. Boik (2007) note that the common principal components model has been employed in many fields such as the genetics, climatology, ontogeny, and other fields (Arnold and Phillips, 1999; Klingenberg et al., 1996; Kulkarni and Rao, 2000; Krzanowski, 1990; Sengupta and Boyle, 1998; Oksanen and Huttunen, 1989).

The Flury-Gautschi (FG) algorithm (Flury and Gautschi, 1986) is the most popular algorithm to minimize (1). Lefkomtch (2004) report that the FG "is computationally expensive, especially for large and/or many matrices." Boik (2007) agree, pointing out that the Flury-Gautschi algorithm is slow in higher ($p > 5$) dimensions. Browne and McNicholas (2012) also show that the FG algorithm is slow in high dimensions. This limitation in application of the FG algorithm has had knock-on effects in methods that use it. For

*E-mail: rbrowne@uoguelph.ca. Tel: +1-519-824-4120, ext. 53034.

example, in a high dimensional mixture modelling application, Bouveyron et al. (2007) avoid the common principal component models stating that they "require a complex iterative estimation based on the FG algorithm (Flury and Gautschi, 1986) and therefore they will be not considered here." To overcome a slow algorithm in high dimensions, Browne and McNicholas (2012) implemented an accelerated line search (ALS) for optimization on the orthogonal Stiefel manifold in a mixture modelling application and showed that this outperforms the FG method in high dimensions and reduces the number of degenerate solutions produced by the EM algorithm. In their ALS, Browne and McNicholas (2012) do not exploit the convexity of the objective function. In this paper, however, we exploit convexity to obtain several simple majorization-minizmation (MM) algorithms (c.f. Hunter and Lange, 2000; Hunter, 2000) following methodology from Kiers (2002). We then compare all algorithms that minimize (1) in terms of convergence and computational time.

2 Minimization on the orthogonal Stiefel manifold

2.1 Flury Method

Flury and Gautschi (1986) suggest an algorithm based on pairwise minimization of the matrix $\mathbf{D}$. That is, each pair of columns or eigenvectors of $\mathbf{D}$ is updated while holding the others fixed. These updates are based on the eigendecompostion of 2×2 matrices summed across groups. Then we are required to loop through all pairs of columns of the matrix $\mathbf{D}$ to complete a single iteration. This makes the Flury method ineffective in higher dimensions. See Flury and Gautschi (1986) for details on the algorithm.

2.2 Accelerated Line Search

An accelerated line search algorithm (ALS) on a manifold consists of selecting a search direction in the tangent space and then moving this direction until a 'reasonable' decrease in the objective function is found. Browne and McNicholas (2012) introduces an ALS algorithm to minimize the function in equation (1). An extensive review of optimization on matrix manifolds is given by Absil et al. (2008). This methods requires tunning parameters and we use the values suggested by Browne and McNicholas (2012).

2.3 MM Algorithm 1

We can exploit the convexity of the objective function (1) to obtain a MM algorithm similar that given in Kiers (2002). Three different MM algorithms are presented and each algorithm has a surrogate function of the form

$$f(\mathbf{D}) = \sum_{g=1}^{G} \text{tr}\{\mathbf{W}_g \mathbf{D} \mathbf{A}_g^{-1} \mathbf{D}'\} \leq C + \text{tr}\{\mathbf{F}_t \mathbf{D}\}$$

where C is a constant that does not depend on $\mathbf{D}$, $\mathbf{F}_t = \sum_{g=1}^{G} \left(\mathbf{A}_g^{-1} \mathbf{D}_t' \mathbf{W}_g - \omega_k \mathbf{A}_g^{-1} \mathbf{D}_t'\right)$, ω_g is largest eigenvalue of the matrix $\mathbf{W}_g$, and subscript t denotes iteration number. The largest eigenvalue of a matrix can be determined using the power iteration method (von Mises and Pollaczek-Geiringer, 1929). If we obtain the singular value decomposition $\mathbf{F}_t = \mathbf{P}_t \mathbf{B}_t \mathbf{R}_t'$, in which $\mathbf{P}_t$ and $\mathbf{R}_t$ are orthonormal, and $\mathbf{B}_t$ is diagonal, containing the singular values of $\mathbf{F}_t$ on the diagonal. Then the update of the matrix $\mathbf{D}$ becomes $\mathbf{D}_{t+1} = \mathbf{R}_t \mathbf{P}_t'$ Then we iteratively repeats this process until convergence.

2.4 MM Algorithm 2

In the second MM algorithm, $\mathbf{F}_t = \sum_{g=1}^{G} \left(\mathbf{W}_g \mathbf{D}_t \mathbf{A}_g^{-1} - \alpha_k \mathbf{W}_g \mathbf{D}_t\right)$, where α_g is largest eigenvalue of the matrix $\mathbf{A}_g^{-1}$. Because $\mathbf{A}_g$ is diagonal and positive definite, the largest eigenvalue of $\mathbf{A}_g^{-1}$ is easily determined. The minimum of the surrogate is found using the same method as in Section 2.3.

2.5 MM Algorithm 3

In the third MM algorithm, $\mathbf{F}_t = \sum_{g=1}^{G} \left(\mathbf{W}_g \mathbf{D}_t \mathbf{A}_g^{-1} - \lambda_g \mathbf{D}_t \right)$, where λ_g is largest eigenvalue of the matrix $\mathbf{A}_g^{-1} \otimes \mathbf{W}_g$. Because the matrix $\mathbf{A}_g^{-1} \otimes \mathbf{W}_g$ is the Kronecker product of two matrices, we have $\lambda_g = \alpha_g \omega_g$. The minimum of the surrogate is found using the same method as in Section 2.3.

2.6 MM Algorithm 4 for the EVE model

We iterate over MM algorithm 1 and MM algorithm 2.

3 Simulation Study

We simulate various instances of the problem of minimizing (1) to compare our approach to the Flury method and the accelerated line search used in Browne and McNicholas (2012). We randomly generated $\mathbf{W}_1, \ldots, \mathbf{W}_G$ where each was produced from a $p+1$ observations from the p-dimensional standard normal distribution. In addition, we randomly generated the diagonal elements $\mathbf{A}_1, \ldots, \mathbf{A}_G$ from the half-normal distribution. Then we varied the number of dimensions p. We used the identity matrix as a starting value for each algorithm and then we ran until convergence. For each simulation, we recorded the system time, the number of iterations, and value of the objective function at the converged solution.

Table 1 shows averages of the system times and the number of iterations from 100 simulations of the six algorithms. The table also gives the the relative difference between the minimum and the converged minimum ('% Diff.') for each case. For a particular simulation, if $\{t_1, \ldots, t_6\}$ are the values of the objective function from the converged solutions from the six algorithms and we let $t_{\min} = \min\{t_1, \ldots, t_6\}$, then difference percentage for algorithm k is $(t_k - t_{\min})/t_{\min}$, for $k = 1, \ldots 6$. Note that if an algorithm has a large '% Diff.' then we could use a stricter convergence criteria to improve the result. However, a stricter convergence criteria will also increase the number of iterations and thus the system time. To facilitate comparison, we used the same convergence criteria for each algorithm.

Table 1: The average system times (in seconds), iterations and difference between convergence value and minimum from the six algorithms.

	p = 5, G = 5			p = 20, G = 5		
Method	Time	Iter.	% Diff.	Time	Iter.	% Diff.
ALS	0.038	34	0.050	0.292	83	8.549
Flury	0.050	10	0.011	4.016	27	0.016
MM 1	0.055	60	0.007	0.323	218	0.362
MM 2	0.027	85	0.010	0.235	318	1.045
MM 3	0.165	179	0.017	0.872	580	2.526
MM 4	0.035	32	0.001	0.263	128	0.180

Table 1 illustrates that the Flury method becomes computationally infeasible when the dimension increases from five to twenty. Also, it seems the ALS method and MM 3 algorithm tend to converge prematurely. Conversely, MM 4 seems to retain computational efficiency while maintaing the same convergence rate as the Flury algorithm. Table 2 tells the same story as Table 1 but with higher dimensions. Results for the Flury method are not reported in the right-hand column of Table 2 due to prohibitive computational time. In the left-hand column of Table 2, the Flury method converged to the smallest value in each simulation. The MM 4 gives similar results to the Flurry algorithm but is just as fast as the ALS algorithm. We note that the running parameters for ALS algorithm could be adjusted to optimize the performance of the ALS algorithm.

Table 2: The average system times (in seconds), iterations and difference between convergence value and minimum from the six algorithms.

	p = 50, G = 5			p = 100, G = 5		
Method	Time	Iter.	% Diff.	Time	Iter.	% Diff.
ALS	2.85	174	0.347	9.66	78	0.500
Flury	101.71	33	0.000			
MM 1	2.22	303	0.025	10.51	290	0.010
MM 2	2.76	565	0.080	22.12	836	0.120
MM 3	8.62	961	0.187	32.85	1521	0.355
MM 4	2.26	214	0.016	12.78	230	0.000

4 Discussion

We find that a MM algorithm is just as fast as the ALS algorithm introduced by Browne and McNicholas (2012) but has the same properties as the Flury method (Flury and Gautschi, 1986). In addition, the MM algorithm does not have any tunning parameters unlike the ALS algorithm. This will allow the implementation of techniques for higher dimensional problems for which the Flury method is too slow. Examples include the method of Bouveyron et al. (2007) for clustering high dimensional data and parameter estimation for some of the mixture models considered by Celeux and Govaert (1995).

References

Absil, P.-A., Mahony, R., Sepulchre, R., 2008. Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ.

Arnold, S., Phillips, P., 1999. Hierarchical comparison of genetic variance-covariance matrices. II. coastal-inland divergence in the garter snake, thamnophis elegans. Evolution 53, 1516–1527.

Boik, R. J., 2003. Principal component models for correlation matricess. Biometrika 90, 679–701.

Boik, R. J., 2007. Spectral models for covariance matrics. Biometrika 89, 159–182.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. Computational Statistics and Data Analysis 52, 502–519.

Browne, R., McNicholas, P., 2012. Orthogonal stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. Statistics and ComputingIn review follow revisions.

Celeux, G., Govaert, G., 1995. Gaussian parsimonious clustering models. Pattern Recognition 28 (5), 781–793.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39 (1), 1–38.

Flury, B. W., Gautschi, W., 1984. Common principal components in k groups. Journal of the American Statistical Association 79 (388), 892–898.

Flury, B. W., Gautschi, W., 1986. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. Journal on Scientific and Statistical Computing 7 (1), 169–184.

Hunter, D., 2000. MM algorithms for generalized Bradley-Terry models. Annals of Statistics 32, 386–408.

Hunter, D., Lange, K., 2000. Quantile regression via an MM algorithm. Computational Statistics and Data Analysis 9, 60–77.

Kiers, H., 2002. Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. Computational Statistics and Data Analysis 41, 157–170.

Klingenberg, C., Neuenschwander, B., Flury, B., 1996. Ontogeny and individual variation: Analysis of patterned covariance matrices with common principal components. Systematic Biology 45, 135–150.

Krzanowski, W. J., 1990. Between-group analysis with heterogeneous covariance. matrices: The common principal component model. Journal of Classification 7, 81–98.

Kulkarni, B., Rao, G., 2000. The common principal components approach for clustering under multiple sampling. J. Indian Soc. Agric. Statist. 53, 1–11.

Lefkomtch, L. P., 2004. Consensus principal components. Biometrical Journal 35, 567–580.

Merbouha, A., Mkhadri, A., 2004. Regularization of the location model in discrimination with mixed discrete and continuous variables. Computational Statistics and Data Analysis 45, 463–576.

Oksanen, J., Huttunen, P., 1989. Finding a common ordination for several data sets by individual differences scaling. Plant Ecology 83, 137–145.

Schott, J., 1998. Estimating correlation matrices that have common eigenvectors. Computational Statistics and Data Analysis 27, 445–459.

Sengupta, S., Boyle, J., 1998. Using common principal components for comparing GCM simulations. Journal of Climate 11, 816–830.

von Mises, R., Pollaczek-Geiringer, H., 1929. Praktische verfahren der gleichungsauflösung . Zeitschrift für Angewandte Mathematik und Mechanik 9 (1), 58–77.

Yang, K., Shahabi, C., 2006. An efficient k nearest neighbor search for multivariate time series. Information and Computation 205, 65–98.